KL Divergence and MLE

Tri Nguyen

nguyetr9@oregonstate.edu

August 17, 2022

Claim. Maximizing likelihood is equivalent to minimizing KL divergence between 2 distributions x and x.

Derivation. Assume we have n i.i.d samples $x_1, \ldots, x_n$ drawn from unknown distribution P. We wish to find a parameter $\theta(P)$ of P using an estimator $\widehat{\theta}(x_1, \ldots, x_n)$.

To see the connection, let's write down the KL divergence

$$\begin{aligned} D_{\text{kl}}(X|\theta \parallel X|\widehat{\theta}) &= \mathbb{E}_{X\sim P(\cdot;\theta)} \left[\log \frac{P(X;\theta)}{P(X;\widehat{\theta})} \right] \\ &= \mathbb{E}_{X\sim P(\cdot;\theta)} \left[\log P(X;\theta) \right] - \mathbb{E}_{X\sim P(\cdot;\theta)} [\log P(X;\widehat{\theta})] \end{aligned}$$

Hence,

$$\arg\min_{\widehat{\theta}} \ D_{\text{kl}}(X|\theta \parallel X|\widehat{\theta}) = \arg\max_{\widehat{\theta}} \ \mathbb{E}_{X\sim P(\cdot;\theta)} \left[\log P(X;\widehat{\theta}) \right]$$

Obviously we cannot evaluate expectation on the RHS since we do not know θ. However, we have an approximation of this term thanks to n i.i.d samples $x_1, \ldots, x_n$ which are drawn from this exact distribution $P(\cdot;\theta)$. And this is nothing but the MLE recipe:

$$\begin{aligned} \arg\max_{\widehat{\theta}} \ \log P(x_1, \ldots, x_n; \widehat{\theta}) &= \arg\max_{\widehat{\theta}} \ \sum_{i=1}^{n} \log P(x_i; \widehat{\theta}) \\ &= \arg\max_{\widehat{\theta}} \ \frac{1}{n} \sum_{i=1}^{n} \log P(x_i; \widehat{\theta}) \\ &\approx \arg\max_{\widehat{\theta}} \ \mathbb{E}_{X\sim P(\cdot;\theta)} \log P(X; \widehat{\theta}) \end{aligned}$$